LIVING AT THE EXTREME EDGE:
TOWARDS DISTRIBUTED HIGH PERFORMANCE
ARTIFICIAL INTELLIGENCE

RUDY LAUWEREINS – VICE PRESIDENT IMEC & PROFESSOR KU LEUVEN
THINK DISTRIBUTED…

- Storage & Processing €, GB
- Bandwidth €, Mbps
- Latency (us)
- Energy
- Utility
- Privacy
- Security: Confidentiality, Integrity, Availability
- Reliability
- Adaptive learning
- Joint learning
- Extreme Edge
5 FOCUS AREAS

- Storage & Processing €, GB
- Bandwidth €, Mbps
- Latency (us)

Communication

Edge

Extreme Edge

Distributed Trust

Security: Confidentiality, Integrity, Availability

Reliability

Joint Learning

Adaptive learning
Joint learning

Sensor Fusion
A CASE DRIVEN APPROACH

- Use Case Definition
- Critical Requirements
- Solution Architecture
- Use Case Roadmap
- Consolidated Roadmap
# SUMMARY OF DISTRIBUTED AI USE CASES AND REQUIREMENTS

<table>
<thead>
<tr>
<th>DESCRIPTION</th>
<th>Security</th>
<th>Mobility</th>
<th>Sequencing</th>
<th>AR</th>
<th>Industrial IoT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Biometric authentication using gait</td>
<td></td>
<td>Supervised intersections</td>
<td>Sequencing of cancer cells in the blood</td>
<td>From single user to real-time, collaborative and perceptive AR</td>
<td>Smart factory quickly reconfiguring (c/r)obot production line</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>STORAGE/MEMORY</th>
<th>PROCESSING</th>
<th>COMMUNICATION</th>
<th>LATENCY</th>
<th>POWER CONSUMPTION</th>
<th>PRIVACY</th>
<th>SECURITY</th>
<th>RELIABILITY &amp; AVAILABILITY</th>
<th>ADAPTIVE LEARNING</th>
<th>JOINT LEARNING</th>
</tr>
</thead>
</table>
ADDITIONAL USE CASES
DISTRIBUTED & COLLABORATIVE AI

Collaborative Drones

Precision agriculture & agri-food

Smart City Digital Twins
CASE STUDY: GENERIC SCREENING PLATFORM FOR BLOOD OR WATER

From drop of blood to high-quality analysis within 1 hour

- Blood test
- Cell Sorting
- Sequencing

Results

- Virus, Cancer, RNA
  - 60 min
- 10M DNA fragments, 800K fps
  - 2 TB/s
- < $300
  - 80 TOP/s

Generic Screening Platform for Blood or Water

- Match against Virus Fingerprints, Cancer Fingerprints, RNA Sequencing, Chronical Oncology Follow-up
- Water: Match against Virus Fingerprints
- Battery powered devices needed for field operations in e.g., emergency care (blood), field operations (blood and water), the latter specifically in developing countries and for military use.
HIGH ENERGY EFFICIENCY NEURAL NETWORK HARDWARE BENCHMARK – DR. EUGENE TARGETS

-sourced from https://nicsefc.ee.tsinghua.edu.cn/projects/neural-network-accelerator/
# DHPAI PROGRAM

## HEALTH – DR. EUGENE USE CASES

<table>
<thead>
<tr>
<th>1-2 years</th>
<th>2-5 years</th>
<th>5+ years</th>
</tr>
</thead>
</table>
| **Combining sensors**  
Training models end-to-end fusing multiple sensor inputs on single device | **Distributed sensors**  
Sensor information coming from devices distributed in the network | **High performance A.I.**  
Optimized A.I. models for optimized hardware |
| **Robust sensor fusion**  
Robust against sensors added / removed / failing | **Transfer learning**  
Transfer learned models to a new instance of the system:  
- new environment  
- other sensors on different locations  
- different agents  
- … | **World modeling**  
A.I. agents that understand the world they operate in:  
- process multiple sensors inputs  
- enable to plan ahead  
- multi-agent interactions  
- zero-shot learning through “imagination”  
- … |

**Sensor Fusion**

**Joint Learning**

### Applications

- Virus Scanner
- Oncology Scanner
- Field Updateable Scanners
- Field Adaptive Scanners
- Oncology Treatment Decision Support
MOBILITY: SUPER-VISION

From autonomous traffic to supervised intersections

Smart intersections that evolve from stupid ones (monitoring) to smart ones (data exchange between actors) to supervised ones (intersection controls the cars when they enter its domain).

Model scaling implies gradually growing amount of “smartness” i.e. data captured and processed by the traffic.

Neither on device, nor edge or cloud provides a closed solution in our first-order analysis.

Open questions remain on the UI/UX aspects of interactions with bikes, pedestrians (unless we just move them to dedicated lanes and bridges).

- Data streaming easily tops 1Gbps
- Supervised navigation of the intersection requires latencies <10ms
- Realtime awareness requires latencies <20ms
- Significant work needed to define interchange formats
- Data streaming can dominate power budget

- Up to 100 actors
- Data streaming
- Low power
- High power
- ASIC
- FPGA
- GPU
- Neuromorphic
- Comms.
- Trade-off
- IoT
- Mobile
- Automotive
- Cloud
SUPER-VISION
ARCHITECTURE DEFINITION

Notes:
• Bikes and pedestrians can be smartphone-enabled. Bikes can be e-bikes
• Vehicles that are smartphone-enabled can have several levels of integration.
• Sensing vehicles (L2+) have multiple cameras and radars
• Local edge connects wirelessly
• Edge-datacenter is wired
ARCHITECTURE OPTIONS
DEPENDING ON LOCAL PROCESSING AND AWARENESS

Option 1: Supercentral
- All raw data flows to the datacenter

Option 1.5: Smartish cars
- Smart vehicles/infrastructure that can abstract do so
- All data to datacenter

Option 2: Smart cars
- All Smart vehicles/infrastructure abstract
- All data to datacenter

Option 3: Edge
- All raw data flows to the edge

Option 4: Smart cars
- All streams go to the edge
- All smart vehicles/infrastructure abstract
- All data to datacenter

Option X: Fusion/Joint
- Partially processed streams
- Everything to edge

Option 5: IC-what-UC
- Only V2V/V2I communications
- Sharing raw video feeds

Option 5.5: Semi-autobots
- Only V2V/V2I communications
- Smart vehicles/infrastructure that can abstract do so

Research challenge
TRADING OFF PROCESSING WITH TRANSMISSION (QUAD*)
BENCHMARK – TRANSMISSION OF IMAGE DATA VS PROCESSING

Power requirement only to transmit the raw data (lowest latency), where we’re using an (optimistic) $10^{-8}$ J/bit average number.

As can be seen, moving all processing to a central cloud can consume almost as much energy as the processing itself.

Compression is equivalent to (partial) local processing with one key difference: latency. ●, ○ are factor 1000 (HEVC target), 100

*QUAD = quick-and-dirty
Closed-loop sensor-based control for robotic manipulators

- Closed-loop robot control requires low latency processing (~10ms) of high resolution sensors (i.e. camera, LIDAR, ...).
- Currently, processing one sensor on a high-end GPU is feasible, but multiple sensors create an I/O bottleneck.
- Neuromorphic hardware stacked on the sensing/actuating chip (“extreme edge”) could help to:
  - Reduce the latency and bandwidth to the fusing A.I.
  - Scale number of sensors as each sensor has dedicated processing power
  - Increase the number of actuators enabling joint learning
# Latency Requirements for Industry 4.0

**Source:** 3GPP 22.804

<table>
<thead>
<tr>
<th>Use case (high level)</th>
<th>Availability</th>
<th>Cycle time</th>
<th>Typical payload size</th>
<th># of devices</th>
<th>Typical service area</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Motion control</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Printing machine</td>
<td>&gt;99.9999%</td>
<td>&lt; 2 ms</td>
<td>20 bytes</td>
<td>&gt;100</td>
<td>100 m x 100 m x 30 m</td>
</tr>
<tr>
<td>Machine tool</td>
<td>&gt;99.9999%</td>
<td>&lt; 0.5 ms</td>
<td>50 bytes</td>
<td>~20</td>
<td>15 m x 15 m x 3 m</td>
</tr>
<tr>
<td>Packaging machine</td>
<td>&gt;99.9999%</td>
<td>&lt; 1 ms</td>
<td>40 bytes</td>
<td>~50</td>
<td>10 m x 5 m x 3 m</td>
</tr>
<tr>
<td><strong>Mobile robots</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cooperative motion control</td>
<td>&gt;99.9999%</td>
<td>1 ms</td>
<td>40-250 bytes</td>
<td>100</td>
<td>&lt; 1 km²</td>
</tr>
<tr>
<td>Video-operated remote control</td>
<td>&gt;99.9999%</td>
<td>10 – 100 ms</td>
<td>15 – 150 kbytes</td>
<td>100</td>
<td>&lt; 1 km²</td>
</tr>
<tr>
<td><strong>Mobile control panels with safety functions</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Assembly robots or milling machines</td>
<td>&gt;99.9999%</td>
<td>4-8 ms</td>
<td>40-250 bytes</td>
<td>4</td>
<td>10 m x 10 m</td>
</tr>
<tr>
<td>Mobile cranes</td>
<td>&gt;99.9999%</td>
<td>12 ms</td>
<td>40-250 bytes</td>
<td>2</td>
<td>40 m x 60 m</td>
</tr>
<tr>
<td><strong>Process automation (process monitoring)</strong></td>
<td>&gt;99.99%</td>
<td>&gt; 50 ms</td>
<td>Varies</td>
<td>10000</td>
<td>devices per km²</td>
</tr>
</tbody>
</table>
HIGH ENERGY EFFICIENCY NEURAL NETWORK HARDWARE
BENCHMARK – PROCESSING IMAGES OF ONE SINGLE CAMERA

based on https://nicsefc.ee.tsinghua.edu.cn/projects/neural-network-accelerator/
DHPAI PROGRAM
INDUSTRIAL IOT – ROBOTICS USE CASES

1-2 years

**Combining sensors**
Training models end-to-end fusing multiple sensor inputs on single device

**Distributed sensors**
Sensor information coming from devices distributed in the network

**Robust sensor fusion**
Robust against sensors added / removed / failing

2-5 years

**Combining data sources between learning agents**
- Learn locally on each device using it’s own data source
- Improve by sharing information, without sharing the whole local dataset

**Robust joint learning**
Robust against agents being added/removed to the system

5+ years

**Distributed sensors**
Sensor information coming from devices distributed in the network

**Transfer learning**
Transfer learned models to a new instance of the system:
- new environment
- other sensors on different locations
- different agents
- …

**Robust joint learning**
Robust against agents being added/removed to the system

**World modeling**
A.I. agents that understand the world they operate in:
- process multiple sensors inputs
- enable to plan ahead
- multi-agent interactions
- zero-shot learning through "imagination"
- …

**High performance A.I.**
Optimized A.I. models for optimized hardware

Applications

- **Multi-camera quality inspection on assembly line**
- **Multi-sensor workcell tracking**
- **Collaborative drones**
- **Flexible assembly line (re)configuration**
- **Real-time, closed-loop robot control**
embracing a better life